Weekly Report

***Time**: 07/09/2012 – 07/15/2012*

This week, I was working on implementation of MP uncertain objects. So far, the probability estimation part is done. And I am working on implementing similarity estimation with relative entropy.
The following is an initial draft of our approach.

Multidimensional Projection of Uncertain Datasets

This section describes the main concept of our approach. First, we treat each uncertain object as a random variable following a probability distribution. Then, the KL divergence i.e. relative entropy, which measures the similarity between tow distributions, is estimated between each pair of random variables. Finally, an MDS algorithm is employed.

Uncertain object and probability distribution

In this paper, each uncertain object is regarded as a random variable with a probability distribution in a domain D. Typically, the exact probability distribution of each random variable is unknown beforehand. Instead, it is derived from a set of observations.
As we know, kernel density estimation (KDE) is a non-parametric method to estimate the probability for a random variable. Given a series of observations of a random variable, a kernel such as Gaussian is associated to each observation. And the density at a given position is the sum of influence from all kernels.
In this paper, Gaussian kernel is adopted. Finally, the d-dimensional kernel density estimator is defined as

$$P(x) = \frac{1}{|P|(\sqrt{2\pi}h)^d} \sum_{p \in P} e^{\frac{(x-p)^2}{h^2}}$$

where h is the bandwidth witch determines the influence region of each kernel and P is a set of d-dimensional samples.

Similarity estimation

In general, the Euclidian distance is calculated as a measurement for any pair of certain objects. This simple strategy is preferred in a large number of multidimensional projection approaches such as IsoMap, LLE. However, there is no such direct existing distance for uncertain object. A naive way to measure the similarity of a pair of uncertain objects is to replace each uncertain object with the mean value of all observations and use the Euclidian distance to approximate the similarity among uncertain object. Unfortunately, this would lose much information with the only mean value and give rise to wrong projection result.
In the filed of information theory, KL divergence is a prevailing metric to quantify the similarity of the given tow distributions. Given tow probability distribution functions f and g, the KL divergence is defined as

$$D(f|g) = \int_D f(x) \log \frac{f(x)}{g(x)} d_x$$

The KL divergence is defined only in the case where for any x in domain D if f(x)>0 then g(x)>0. Conventionally, $0 log \frac{0}{p}$ for any $p \neq 0$ is defined as 0. The base of log is 2. Note that, KL divergence is not symmetric, that is, $D(f|g) \neq D(g|f)$ with the exception that $f = g$.

In this paper, given tow uncertain object P and Q, we calculate the KL divergence of their corresponding probability distribution function as the similarity of P and Q.

$$S(P, Q) = D(P(x)|Q(x))$$

Here, P(x) and Q(x) are the probability distribution function of P and Q.

Accelerating probability estimation

Miscellaneous

Work to do in next week

- Finishing implementing similarity calculation of uncertain objects
- Keep on drafting our multidimensional projection approach

Reference:

[1] Clustering Uncertain Data Based on Probability Distribution Similarity.